ELG Pilot Project: Textual paraphrase dataset for deep language modelling
Textual paraphrase dataset for deep language modelling
Paraphrase: “Two statements having the same contextual meaning, using a different wording”

“Calls from one EU country to another will have a price limit of 19 cents”
“EU-internal calls will cost no more than 19 cents”

- Deep language models should be able to model meaning and **encode mutual paraphrases into similar representations**
- **Paraphrase is a natural task for NLU model training**
- Human judgements needed to identify the truly **interesting cases which have little lexical overlap**
<table>
<thead>
<tr>
<th>Class</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>4&gt;</td>
<td>-</td>
</tr>
</tbody>
</table>

**Carter... Mitä tapahtuisi jos vain irrottaisimme töpselin?**

**Carter mitä tapahtuisi, jos vain... vedämme johdon irti?**

As identified in the source document

Rewritten to a perfect paraphrase (where possible)
Primary outcome

- The primary outcome is a large Finnish paraphrase dataset
  - Pairs identified in text
  - Their context
  - Class
  - Rewritten version where applicable
- 100,000+ pairs is the target, model training - finetuning - testing
- Open license, data release deadline 31.5.2021